ANALYSIS OF ARTIFICIAL INTELLIGENCE AND DATA ACT BASED ON ETHICAL FRAMEWORKS

Sun Gyoo Kang[*]

Abstract: Canada has recently introduced the Digital Charter Implementation Act of 2022. Both at the federal level as well as at the provincial level, different governments are trying to move forward with emerging technology by amending and implementing laws and regulations to put a safety net against various emerging risks, but at the same time, promote the growth of the innovative industry in Canada. Indeed, as one of the first countries, Canada introduced legislation regarding artificial intelligence (Artificial Intelligence and Data Act), and the goal of this paper is to check if the new legislation would encompass the basic artificial intelligence ethical frameworks such as privacy, accountability, transparency/explainability, fairness, and safety & security, which were recommended in three (3) different ethical declarations on artificial intelligence, among many others. The Artificial Intelligence and Data Act contains important sections incorporating accountability, transparency/explainability, fairness, and safety & security. Considering a risk-based approach targets high-impact systems and requires risk management and monitoring against the risk of harm and biased output. It also requires persons responsible for high-impact systems to communicate all the necessary information publicly and to be audited in case of concern with its artificial intelligence systems. The new legislation imposes administrative monetary penalties and offenses similar to the proposed regulations in the EU. As for privacy, similar to the EU, it is outside of the new law’s scope, but another law covers the topic independently already Nevertheless, the Artificial Intelligence and Data Act also has some pitfalls. It lacks clarity and specific requirements found in the law of the EU. Furthermore, the scope is an issue as it only covers private sector actors, and there is doubt about the real independence and neutrality of the commissioner.

Keywords: Artificial Intelligence; Artificial Intelligence and Data Act; Risk-based Approach; Canada; EU

[*] National Bank of Canada, Canada.

Table of Content

INTRODUCTION

On June 16, 2022, the Federal Government of Canada proposed Bill C-27[1], which made the news for its newly amended privacy law, the Consumer Privacy Protection Act ("CPPA"). In addition to the newly amended federal privacy law, the Artificial Intelligence and Data Act ("AIDA") was introduced and has created a wave of questions among the artificial intelligence industry. One of the hot questions is on what exactly is a "*high-impact system*"[2]. The answers to many of the questions asked by the designers, data scientists and engineers will likely be included in the regulations that will be subsequently introduced and without the regulations, it is hard to evaluate and compare with other jurisdictions (i.e.: EU's Artificial Intelligence Act).

As a matter of fact, this paper does not aim to compare this new legislation with other proposed legislations in the EU or the United States. Also, the regulations that will support the proposed legislation are yet to come out so this paper will try to analyse just with the proposed law itself.

Nonetheless, the infinite possibility of what an artificial intelligence system can do is not easily measurable. As advocated in the Asilomar Principle[3]human beings cannot confirm the upper limits of what an artificial intelligence can do. And we know that law has its own limit as most of the time, the laws and regulations would rather react and will not cover all the aspects in our life. For that reason, ethics then naturally becomes an important aspect for artificial intelligence.

In order to perform a thorough analysis, three (3) ethical principle guidelines were analyzed: Montréal Declaration Responsible AI[4], The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems[5] and Asilomar AI Principle[6]. Then the five (5) main ethical frameworks were chosen: 1) Privacy, 2) Fairness, 3) Accountability, 4) Transparency/ Explainability and 5) Safety and Security for analysis. At the end, the main articles of the AIDA will be analyzed based on these ethical frameworks to see if AIDA is commensurate with the artificial intelligence ethical framework principles.

[1] Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts (Canada)*, 1st session, 44th Parliament, 2022 (Consumer Privacy Protection Act)).

[2] Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts (Canada)*, 1st session, 44th Parliament, 2022, section 5 (1) of the Artificial Intelligence and Data Act (Canada) (Artificial Intelligence and Data Act).

[3] Principle 19) Capability Caution of the ASILOMAR AI PRINCIPLES <https://futureoflife.org/2017/08/11/ai-principles/> accessed on 8 August 2022 (Asilomar AI Principles).

[4] Montréal Declaration Responsible AI <www.montrealdeclaration-responsibleai.com/_files/ugd/ebc3a3_506ea08298cd4f8196635545a16b071d.pdf> accessed on 8 August 2022.

[5] The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems <www.torontodeclaration.org/wp-content/uploads/2019/12/Toronto_Declaration_English.pdf> accessed on 8 August 2022 (Toront Declaration).

[6] Asilomar AI Princples <https://futureoflife.org/2017/08/11/ai-principles> accessed on 8 August 2022.

I. ARTIFICIAL INTELLIGENCE AND DATA ACT

A. Scope

Unlike EU's Artificial Intelligence Act ("AIA")[7], AIDA only applies to private sectors and does not apply to the public sector. It would be interesting to know the exact reason why the public sector is out of scope but if we compare with the privacy legislation at the federal level in Canada, it might be logical that the legislation related to artificial intelligence also follows the same structure. In the federal privacy legislation, there is the PIPEDA[8] that is applicable to the private sector and Privacy Act[9] that is applicable to the public sector. Another guess is that the public sector already has the Directive on Automated Decision-Making ("Directive") so the federal government might have thought that the Directive was enough for the moment. Nonetheless, there are some concerns[10] with the narrow scope of the AIDA, and the fact that the public sector is not in scope could be a major issue. Also, the Directive has its own issues as it does not cover areas impacting federal employees including the hiring process[11].

In the United States, after a failed attempt in 2019 to introduce a bill on automated-decision making, the Algorithmic Accountability Act of 2022[12] ("AAA") was reintroduced in 2022 with some amendments from the 2019 version. Similar to AIDA, AAA, as the current form, would be applicable only to private sector actors as well. Indeed, AAA describes them as covered entities[13] and captures two types of businesses: 1) big firms deploying augmented critical decision processes ("ACDP")[14] and 2) medium size firms deploying automated decision-making systems ("ADS") which will be used by the big firms[15].

B. Definition

1. Artificial Intelligence System

As of today, there is no single set of definition of what an artificial intelligence system is but section 2 of the AIDA tries to define it as follows:

> *artificial intelligence system means a technological system that, autonomously or partly autonomously, processes data related to human activities through the use of a genetic algorithm, a neural network, machine learning or another*

[7] Commission, 'Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts' COM/2021/206 final (Artificial Intelligence Act).

[8] *Personal Information Protection and Electronic Documents Act,* S.C. 2000, c. 5.

[9] *Privacy Act*, R.S.C., 1985, c. P-21.

[10] 'Roundtable on the Artificial Intelligence and Data Act' (Centre for Media, Technology & Democracy, July 12 2022), <https://youtu.be/Ll46lPnfvZU> accessed on 8 August 2022.

[11] Omar Bitar, Benoit Deshaies & Dawn Hall, '3rd Review of the Treasury Board Directive on Automated Decision-Making' (2022), <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4087546> access on 8 august 2022.

[12] H.R.6580 - 117th (2021-2022): Congress Algorithmic Accountability Act of 2022, (2022, February 3). https://www.congress.gov/bill/117th-congress/house-bill/6580 (Algorithmic Accountability Act).

[13] Artificial Intelligence Act, Section 2(7)(a).

[14] Artificial Intelligence Act, Section 2(7)(a)(i).

[15] Artificial Intelligence Act, Section 2(7)(a)(ii).

> *technique in order to generate content or make decisions, recommendations or predictions. (système d'intelligence artificielle)*

So it must basically be an autonomous or partly autonomous system which processes data and the data must be related to human activities.To add, the definition provides examples of techniques such as genetic algorithm, a neural network, machine learning and any other technique to generate content or make decisions, recommendations or predictions. Obviously, the last part is to be future-proof. Furthermore, *human activities* are not defined in AIDA.

The EU's AIA's definition of the artificial intelligence system is as follows:

> *'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with*[16]*;*

The EU's AIA definition is more specific by providing a list of technologies in its Annex I, which is amendable as well, and below are what EU considers as artificial intelligence techniques and approaches:

> *(a)Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;*
>
> *(b)Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;*
>
> *(c)Statistical approaches, Bayesian estimation, search and optimization methods*[17]*.*

In the United States, Section 2 (2) of AAA describes automated decision system as below:

> *AUTOMATED DECISION SYSTEM.—The term "automated decision system" means any system, software, or process (including one derived from machine learning, statistics, or other data processing or artificial intelligence techniques and excluding passive computing infrastructure) that uses computation, the result of which serves as a basis for a decision or judgment.*

This definition seems to cover more than the AIDA and AAA as it is broader than just the artificial intelligence systems.

2. Biased Output

Section 5 (1) of the AIDA's definition of a biased output is as follows:

[16] Artificial Intelligence Act, Article 3.

[17] Artificial Intelligence Act, Appendix I.

> *biased output means content that is generated, or a decision, recommendation or prediction that is made, by an artificial intelligence system and that adversely differentiates, directly or indirectly and without justification, in relation to an individual on one or more of the prohibited grounds of discrimination set out in section 3 of the Canadian Human Rights Act, or on a combination of such prohibited grounds. It does not include content, or a decision, recommendation or prediction, the purpose and effect of which are to prevent disadvantages that are likely to be suffered by, or to eliminate or reduce disadvantages that are suffered by, any group of individuals when those disadvantages would be based on or related to the prohibited grounds. (résultat biaisé)*

AIDA seems to focus on the result or output coming from the artificial intelligence system. This definition specifically refers to the discrimination set out in section 3 of Canadian Human Rights Act ("CHRA")[18]. It includes "*race, national or ethnic origin, colour, religion, age, sex, sexual orientation, gender identity or expression, marital status, family status, genetic characteristics, disability and conviction for an offence for which a pardon has been granted or in respect of which a record suspension has been ordered*". Furthermore, there might be a part in the regulation or a stand-alone regulation that may define a biased output[19].

On the other hand, the definition seems to not include "biased output" from artificial intelligence systems that are used to prevent discrimination and harm. This may be a concern as recently FTC published a report noting that even artificial intelligence used to counter bias and discrimination may bring additional harms such as:

> *Inherent design flaws and inaccuracy: AI detection tools are blunt instruments with built in imprecision and inaccuracy. Their detection capabilities regarding online harms are significantly limited by inherent flaws in their design such as unrepresentative datasets, faulty classifications, failure to identify new phenomena, and lack of context and meaning.*
>
> *Bias and discrimination: In addition to inherent design flaws, AI tools can reflect biases of its developers that lead to faulty and potentially illegal outcomes. The report provides analysis as to why AI tools produce unfair or biased results. It also includes examples of instances in which AI tools resulted in discrimination against protected classes of people or overblocked content in ways that can serve to reduce freedom of expression.*
>
> *Commercial surveillance incentives: AI tools can incentivize and enable invasive commercial surveillance and data extraction practices because these technologies require vast amounts of data to be developed, trained, and used. Moreover, improving AI tools accuracy and performance can lead to more invasive forms of surveillance*[20].

[18] Canadian Human Rights Act, R.S.C., 1985, c. H-6.

[19] Artificial Intelligence and Data Act, Section 36 (a).

[20] Federal Trade Commission, 'Combatting Online Harms Through Innovation' (2022, June 16), <www.ftc.gov/reports/combatting-online-harms-through-innovation> access on August 8 2022.

3. Harm

Section 5(1) of AIDA also defines what harm means exactly in the context of an artificial intelligence system.

> *Harm means*
>
> *(a) physical or psychological harm to an individual;*
>
> *(b) damage to an individual's property; or*
>
> *(c) economic loss to an individual. (préjudice)*

The goal of AIDA basically is to make responsible persons of high-impact artificial intelligence systems have a risk management program. Along with bias, no harm should be produced from artificial intelligence systems and if produced, there should be controls in place to mitigate the inherent risk.

4. High-Impact System

The focus of AIDA is on high-impact systems. Sections 7 and 8 of AIDA says:

> "7 *A person who is responsible for an artificial intelligence system must, in accordance with the regulations, assess whether it is a high-impact system.*"
>
> "8 *A person who is responsible for a high-impact system must, in accordance with the regulations, establish measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system.*"

Basically, AIDA lets the person perform the assessment to see if its artificial intelligence system is a high-impact system or not. Should it be considered to be a high-impact system, it has the obligation to mitigate the risks. However, AIDA does not really provide a definition. Section 5 (1) of AIDA proposes as a definition: "*high-impact system means an artificial intelligence system that meets the criteria for a high-impact system that are established in regulations*." Until the regulations are published, the industry will not know what exactly is a high-impact system.

Nevertheless, deductions could be made from different sources such as the Directive, AIA and the AAA. The former applies to Canadian governments wishing to utilize automated decision-making system, which by definition is broader[21] than the definition of an artificial intelligence system[22] and the second one applies to private and public organizations in regards to artificial intelligence system[23]. The Directive proposes four (4) impact assessment levels:

- *Level I: little to no impact on*
  - *the rights of individuals or communities,*

[21] Government of Canada, Appendix A of the Directive on Automated Decision-Making (2021, April 1) </www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592#appA>.

[22] Artificial Intelligence and Data Act, section 2.

[23] Artificial Intelligence Act, Article 3 (1).

    - *the health or well-being of individuals or communities,*
    - *the economic interests of individuals, entities, or communities,*
    - *the ongoing sustainability of an ecosystem.*
- *Level II: Moderate impact;*
    - *the rights of individuals or communities,*
    - *the health or well-being of individuals or communities,*
    - *the economic interests of individuals, entities, or communities,*
    - *the ongoing sustainability of an ecosystem.*
- *Level III: High impact; and*
    - *the rights of individuals or communities,*
    - *the health or well-being of individuals or communities,*
    - *the economic interests of individuals, entities, or communities,*
    - *the ongoing sustainability of an ecosystem.*
- *Level IV: Very high impact.*
    - *the rights of individuals or communities,*
    - *the health or well-being of individuals or communities,*
    - *the economic interests of individuals, entities, or communities,*
    - *the ongoing sustainability of an ecosystem.*

In order to check the impact level of an automated decision-making system, one must go through the Algorithmic Impact Assessment tool[24], which consists of 48 questions divided into six (6) risk areas: 1) project, 2) system, 3) algorithm, 4) decision, 5) Impact and 6) data and 33 questions of mitigation areas, which can reduce the residual score. At the end, depending on the final score, the impact level is determined.

As for the AIA, it proposes 4 different types of artificial intelligence systems: (i) an unacceptable risk, (ii) a high risk, and (iii) low or minimal risk. Under the AIA,

[24] Government of Canada, 'Algorithmic Impact Assessment tool', <www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html#toc2-1> accessed on August 8 2022.

the artificial intelligence in the following area are considered to be high risk artificial intelligence system[25]:

- *critical infrastructures (e.g. transport), that could put the life and health of citizens at risk;*
- *educational or vocational training, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);*
- *safety components of products (e.g. AI application in robot-assisted surgery);*
- *employment, management of workers and access to self-employment (e.g. CV-sorting software for recruitment procedures);*
- *essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan);*
- *law enforcement that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);*
- *migration, asylum and border control management (e.g. verification of authenticity of travel documents);*
- *administration of justice and democratic processes (e.g. applying the law to a concrete set of facts).*

Lastly, as for AAA, the previous version of it had a specific definition of what a high-risk automated decision system was[26]. However, in the 2022 version, it refers rather to an augmented critical decision process which is an automated decision system that makes critical decisions. The definition of what a critical decision is:

> *The term "critical decision" means a decision or judgment that has any legal, material, or similarly significant effect on a consumer's life relating to access to or the cost, terms, or availability of—*
>
> *(A) education and vocational training, including assessment, accreditation, or certification;*
>
> *(B) employment, workers management, or self-employment;*
>
> *(C) essential utilities, such as electricity, heat, water, internet or telecommunications access, or transportation;*
>
> *(D) family planning, including adoption services or reproductive services;*

[25] European Commission, 'Regulatory framework proposal on artificial intelligence', (Shaping Europe's digital future 2022, September 29), <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#:~:text=AI%20systems%20identified%20as%20high,e.g.%20scoring%20of%20exams)%3B> accessed on September 30 2022.

[26] Algorithmic Accountability Act, Section 2 (7).

> *(E) financial services, including any financial service provided by a mortgage company, mortgage broker, or creditor;*
>
> *(F) healthcare, including mental healthcare, dental, or vision;*
>
> *(G) housing or lodging, including any rental or short-term housing or lodging;*
>
> *(H) legal services, including private arbitration or mediation; or*
>
> *(I) any other service, program, or opportunity decisions about which have a comparably legal, material, or similarly significant effect on a consumer's life as determined by the Commission through rulemaking*[27]*.*

So, using both the Directive, AIA and AAA as reference, one could assume that similar criteria would probably be used for AIDA but only the regulations would confirm it.

II. WHY ETHICAL GUIDELINES MATTERS

Laws and regulations are the rules covering different human activities and there are usually real consequences when one fails to comply with it. Yet, it takes time to adopt a bill of law and amending an existing one may also take several years. Nevertheless, artificial intelligence is an emerging technology just like quantum computing, metaverse, blockchain and cryptocurrency. Law is something that is often reactive and cannot cover all aspects of human activities.

However, numerous ethical guidelines can chaperone an artificial intelligence system from its inception, design, plan, implementation, test, activation and maintenance. Ethical guidelines are more flexible than law and can catch up with the fast emerging industry such as artificial intelligence. In addition, existing laws or future laws on artificial intelligence must be based on ethical frameworks in order to support the ethical principles. That is why ethical guidelines matter a lot in artificial intelligence and so three (3) of the major ethical guiding principles will be briefly analyzed to have a better understanding of what are the ethical issues around artificial intelligence systems.

A. Ethical Guiding Principles

1. Montréal Declaration Responsible AI

Declared on December 8th 2018 by different university scholars, citizens, artificial intelligence experts and professionals, Montréal Declaration Responsible AI (“MDRAI”) is an important piece of ethical guideline and principles for the artificial intelligence industry. The objectives of MDRAI are:

1. Develop an ethical framework for the development and deployment of AI;

2. Guide the digital transition so everyone benefits from this technological revolution;

[27] Algorithmic Accountability Act, Section 2 (8).

3. Open a national and international forum for discussion to collectively achieve equitable, inclusive, and ecologically sustainable AI development[28].

MDRAI includes ten (10) principles: 1) Well-being, 2) respect for autonomy, 3) protection of privacy and intimacy, 4) solidarity, 5) democratic participation, 6) equity, 7) diversity inclusion, 8) prudence, 9) responsibility and 10) sustainable development.

a. Well-being

MDRAI's first principle promotes the well-being of not only human beings but of all sentient beings. This principle encourages artificial intelligence to be beneficial to human beings and sentient beings and be non-malient to sentient beings. Examples of well-being are the beneficial effects to the economy, health, safety, environment, labor, and etc. Artificial intelligence must be used to support human beings to be healthier by providing better cure to illness[29] and prevention at an earlier phase[30]. Artificial intelligence must be used to support the employees so that individuals are not negatively affected by the system[31]. Furthermore, artificial intelligence should be deployed to support sustainable development[32].

Artificial intelligence must improve the life of individuals and further all sentient beings. Indeed, animals are more and more recognized in different jurisdiction such as the United Kingdom where Animal Welfare (Sentience) Act 2022[33] was enacted to acknowledge animals such as a dog as.

b. Respect for Autonomy

MDRAI's second principle is the respect for autonomy. Autonomy of human beings means that an individual can make independent decisions for its own without any pressure or coercion. With biases such as automatisation bias, the tendency for humans to trust automated made-decisions, and complacency bias, the tendency for humans to not supervise automated made-decisions, it is very important for artificial

[28] Montreal Declaration Responsible AI, 'montréal declaration for a responsible development of artificial intelligence 2018' (2018), </www.montrealdeclaration-responsibleai.com/_files/ugd/ebc3a3_506ea08298cd4f8196635545a16b071d.pdf> accessed on August 8 2022.

[29] Aniek F. Markus, Jan A. Kors & Peter R.Rijnbeek, 'The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies' (2020) 113 Journal of Biomedical Informatics <www.sciencedirect.com/science/article/pii/S1532046420302835> accessed on August 8, 2022.

[30] Jonathan P. Rowe & James C. Lester, 'Artificial Intelligence for Personalized Preventive Adolescent Healthcare' (2020), 67/2 Journal of Adolescent Health <www.sciencedirect.com/science/article/pii/S1054139X20300951> accessed on August 8, 2022.

[31] Marguerita Lane & Anne Saint-Martin, 'The impact of Artificial Intelligence on the labour market' (2021) <www.oecd-ilibrary.org/social-issues-migration-health/the-impact-of-artificial-intelligence-on-the-labour-market_7c895724-en> accessed on August 8, 2022.

[32]Tanveer Ahmad, Dongdong Zhang, Chao Huang, Hongcai Zhang, Ningyi Dai, Yonghua Song, Huanxin Chen, 'Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities' (2021) 289 Journal of Cleaner Production <www.sciencedirect.com/science/article/abs/pii/S0959652621000548> accessed on August 8, 2022.

[33] Animal Welfare (Sentience) Act 2022.

intelligence to respect the autonomy of individuals. As an example, pilots may trust the aircraft machines too much, which may cause accidents time to time[34].

c. Protection of Privacy and Intimacy

The third principle is about privacy and intimacy. This principle will be further discussed below but many jurisdictions have already decided to legislate around the use of personal information already. EU's GDPR[35], Canada's PIPEDA[36], California's CCPA[37], Brazil's LGPD[38] and South Africa's POPIA[39] have been enacted and regulated the use of personal information. Indirectly, artificial intelligence systems are affected as personal information must be collected, used and disclosed or retained.

d. Solidarity

The fourth principle of MDRAI is that artificial intelligence must support the collaborative relationship between individuals. This principle emphasizes the importance of human relationships between people and generation. Furthermore, it puts significance in the fact that artificial intelligence should not replace human beings where it is expected to have quality human relationships. The latter could target, as an example, where general artificial intelligence could be used as part of a romance/date application[40].

e. Democratic Participation

The fifth principle covers a variety of ethical frameworks: transparency, explainability, accountability, Intelligibility, justifiability and accessibility. This paper will cover more in detail the principles of transparency, explainability and accountability later in another section but here are the definitions for each of the three (3) ethical principles.

- *Transparency*: Artificial intelligence is known to be non-transparent. The "black box" issue of complex models is the fact that deep learning models self-learn and even when human-in-the-loop control is put in place, it might be difficult for a professional to understand the result and explain the logic to the users[41].

[34] Alexander Freund, 'Boeing crash: Can machines make better decisions than people?' (2019) DW <www.dw.com/en/boeing-crash-can-machines-make-better-decisions-than-people/a-47920904> accessed on August 8 2022.

[35] Council Regulation (EC) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L119/1 (GDPR).

[36] *Personal Information Protection and Electronic Documents Act,* S.C. 2000, c. 5 (PIPEDA).

[37] Section 3, Title 1.81. 5 of the CCPA, added to Part 4 of Division 3 of the California Civil Code. [3] § 1798.185(a)(1)-(2), (4), (7). [4] § 1798.140(c).

[38] Lei Geral de Proteção de Dados Pessoais (LGPD), Lei n° 13.709/2018.

[39] South Africa. 2013. Protection of personal information act 4 of 2013. Available at < www.gov.za/sites/default/files/gcis_document/201409/3706726-11act4of2013protectionofpersonalinforcorrect.pdf> accessed on August 8 2022.

[40] Adrian David Cheok & Emma Yann Zhang, *Human–Robot Intimate Relationships*, Springer 2019.

[41] Yavar Bathaee, 'The artificial intelligence black box and the failure of intent and causation' (2018) 31 Harvard Journal of Law & Technology. <https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Ba thaee.pdf> accessed on August 8, 2022.

Transparency is the general principle and may include explainability, interpretability and auditability.

- *Explainability*: Explainability is the principle which a designer of the artificial intelligence system can explain the output, input and the model to a user[42].
- *Accountability*: Accountability is the principle where designers or developers should be held accountable for the artificial intelligence system. The OECD principle 1.5 says "*AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.*"[43]

f. Equity

The sixth principle of MDRAI promotes a just and equitable society contributed by the development and use of artificial intelligence systems. This principle describes the fact that artificial intelligence must not be designed to reproduce any types of discrimination. To add, it should be noted that equity differs from equality[44] and MDRAI promotes equity and not equality. More will be discussed later under the "fairness" ethical framework.

g. Diversity inclusion

The seventh MDRAI principle is about diversity inclusion. Nowadays, it is not a mythic statement anymore that diversity is actually a strength to society. McKinsey's research on diversity showed that:

- *"Companies in the top quartile for racial and ethnic diversity are 35 percent more likely to have financial returns above their respective national industry medians" and;*
- *"Companies in the top quartile for gender diversity are 15 percent more likely to have financial returns above their respective national industry medians*"[45].

Thus, the development of artificial intelligence must make sure to promote social and cultural diversity and avoid homogenization of the society.

h. Prudence

The next principle of MDRAI promotes the development of artificial intelligence in such a way to avoid adverse consequences. Here, the focus is on the

[42] GiuliaVilone & LucaLongo, 'Notions of explainability and evaluation approaches for explainable artificial intelligence' (2021)76 Information Fusion <www.sciencedirect.com/science/article/pii/S1566253521001093> accessed on August 8, 2022.

[43] OECD.AI, OECD's AI Principles Accountability (Principles 1.5) <https://oecd.ai/en/dashboards/ai-principles/P9> accessed on August 8, 2022.

[44] George Washington University School of Public Health, 'Equity vs. Equality: What's the Difference?' (2020), <https://onlinepublichealth.gwu.edu/resources/equity-vs-equalit/> accessed on August 8, 2022.

[45] Vivian Hunt, Dennis Layton, and Sara Prince, 'Why diversity matters' (2015) (*McKinsey & Company 2 February 2015)* <www.mckinsey.com/business-functions/people-and-organizational-performance/ou r-insights/why-diversity-matters> accessed on August 8, 2022.

maleficent use of artificial intelligence by humanity and the objective is to limit any harmful use and when the artificial intelligence system is already in use, to be able to restrict the access to it and fix it. More will be discussed later in the safety and security section in this paper.

i. Responsibility

The ninth MDRAI principle is about human responsibility. If a harm was caused by an artificial intelligence system, there must be a person responsible for the results and human beings cannot blame the artificial intelligence. This principle includes the expectation of the development of an artificial intelligence that affects a person's life, quality of life, or reputation to be supervised and approved by a human being. More will be discussed later under the accountability section.

j. Sustainable Development

The last principle promotes the use of artificial intelligence so that it does not cause any harm to the environment. The objective is to the environmentally responsible design, development and use of artificial intelligence systems.

2. The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems

The Toronto Declaration ("TD") was published in May 2018 by Amnesty International[46] and AccessNow[47]. The TD focuses on human rights and is both applicable to the public and private sectors. The TD specifically targets machine learning but it is also applicable in general to all artificial intelligence systems. It is structured as 59 paragraphs and can be grouped into 5 themes as follows:

a. Well-being

The first part talks about the importance of acknowledging and respecting international human rights. Governments must promote human rights while private sectors must respect them. Human rights are universally inalienable rights[48]. The moment a person is born, that person will have equal rights as the others, and unless there is a specific circumstance (i.e.: crime), the human rights cannot be separated from the person. However, often human rights get in conflict with commercial interest (capitalist interest)[49] and therefore TD emphasizes on the central role that governments and private sector actors must play with the emerging technology such as machine learning.

Furthermore, it describes the most important ethical framework under human right: "*the right to equality and non-discrimination*", "*preventing discrimination*" and

[46] Amnesty International <www.amnesty.org/en/> accessed on August 8, 2022.

[47] Access Now <www.accessnow.org>/ accessed on August 8, 2022.

[48] The Office of the High Commissioner for Human Rights, 'What are human rights?', <www.ohchr.org/en/what-are-human-rights#:~:text=Human%20rights%20are%20inalienable.,by%20a%20court%20of%20law> accessed on August 8, 2022.

[49] Bianca Carrera Espriu, 'Capitalism's incompatibility with human rights' compliance' (2021) <www.researchgate.net/publication/352765551_CAPITALISM'S_INCOMPATIBILITY_WITH_HUMAN_RIGHTS'_COMPLIANCE> accessed on August 8, 2022.

"*protecting the rights of all individuals and groups: promoting diversity and inclusion*". The first sub-principle explains what is considered a discrimination. Then it highlights the ideal role of the government and companies as gatekeepers of human rights towards emerging technology. The last part describes that in order to support equality and non-discrimination, inclusion, diversity and equity are key components to achieve them.

b. Duties of States: Human Rights Obligations

The second part of the declaration goes more in detail about the role the governments should play with machine learning including cases where private sectors have partnership with the public sector.

State Use of Machine Learning Systems

TD states that governments may use machine learning systems for all kinds of governmental activities for the public and may include "*the exercise and enjoyment of human rights, rule of law, due process, freedom of expression, criminal justice, healthcare, access to social welfare benefits, and housing*."[50] While using emerging technology such as machine learning, the government must abide by international and national human right law. In order to do so, it recommends three (3) steps to take: 1) perform the identification of risk related to emerging technology such as artificial intelligence and execute regular impact assessment[51], 2) be transparent and accountable for the use of emerging technology[52], and 3) enforce oversight by making sure that government officials understand the risks and be responsible for the private sector partners to make sure that they comply with the human rights law[53].

As an example, the use of emerging technology such as surveillance cameras with facial recognition technology by Canadian federal police authority, Royal Canadian Mounted Police ("RCMP:), made the news[54]in 201. Clearview AI was the service provider to RCMP for AI facial-recognition technology. The Office of Privacy Commissioner ("OPC"), out of the three (3) complaints, found two of them being valid and provided recommendations[55] to RCMP, which has accepted[56]. Should RCMP have taken the above mentioned three (3) steps recommended by the TD, different results would have come out.

Promoting Equality

In addition, the government should not act in a reactive way but must proactively eliminate discrimination. Currently, it is impossible to see the limit of the

[50] Toronto Declaration Paragraph 27.

[51] Toronto Declaration Paragraph 31.

[52] Toronto Declaration Paragraph 32.

[53] Toronto Declaration Paragraph 33.

[54] Moira Warburton, 'Canada police broke law with facial recognition software, regulator finds', *Reuters* (10 June 2021) <https://www.reuters.com/article/us-canada-privacy-idCAKCN2DM208> accessed on August 8, 2022.

[55] At the time of writing this paper, the OPC does not have the power to enforce.

[56] Office of the Privacy Commissioner of Canada, 'RCMP contravened the Act by using certain types of non-conviction information for vulnerable sector checks without consent' (29 March 2021) <www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-federal-institutions/2020-21/pa_20210329_rcmp/> accessed on August 8, 2022.

emerging technology and the harm it can cause to the human society thus the role of the governments is to build programs increasing equity, diversity and inclusion[57].

Holding Private Sector Actors to Account

The last and the most important role of the governments is to make sure that private sector actors are accountable. In order to do that, governments must make laws, regulations and rules governing the use of machine learning. Furthermore, there should be a mechanism in place providing remedy to individuals affected or harmed by the emerging technology.

c. Responsibilities of Private Sector Actors: Human Rights Due Diligence

Just like the public sector, the private sector also requires human rights when using emerging technology. TD expects private sector actors to especially perform what they call human right due diligence independent of what the government's obligations are[58]. In order to do so, TD suggests that private sector actors go through three (3) steps: i. Identify potential discriminatory outcomes, ii. Take effective action to prevent and mitigate discrimination and track responses and iii. Be transparent about efforts to identify, prevent and mitigate against discrimination in machine learning systems[59].

The first step is similar to the one recommended to the governments. The developers must perform risk mapping and impact assessment in advance before deployment of machine learning. Then the next step would be to prevent the risk related to machine learning and put controls in place to prevent harms and discrimination. Furthermore, when the risk is too high or the risk is impossible to be mitigated, TD recommends to not deploy the system. The last step is similar to the second step recommended for the states. Private sector actors must be transparent and be accountable.

d. The Right to an Effective Remedy

In the TD, there is an emphasis on both private sector actors and governments to have a mechanism for redress or remedy when machine learning causes harm or discrimination. Similar to the banking industry, the private sectors may establish an internal mechanism to deal with disputes, complaints or issues around the use of artificial intelligence. As for the governments, they should follow the standards of due process, they should be cautious when deploying machine learning in the justice system, set out accountability and provide remedies through laws and regulations to the victims.

e. Conclusion

In the conclusion part, TD highlights again the fact that the advance of emerging technology must not ignore human rights. Furthermore, governments and private sector actors must work together to ensure there are no harms and discriminations caused by the emerging technology such as machine learning to humanity.

[57] Toronto Declaration Paragraph 37.

[58] Toronto Declaration Paragraph 42.

[59] Toronto Declaration Paragraph 44.

3. Asilomar AI Principles

The Asilomar AI Principles were developed in 2017 by Future of Life institute. The Asilomar AI Principles are 23 in total and are divided into three (3) parts: 1) Research Issues, 2) Ethics and Values and 3) Longer-term issues.

In the first part, Asilomar AI Principles declares that when performing artificial intelligence research, the research must be beneficial[60], there should be an exchange between researchers and policy-makers[61], and should avoid racing for results (profits) without considering the safety measures[62].

The second part of the Asilomar AI Principles focuses on ethics and values such as safety, transparency, responsibility and privacy. As for transparency[63], it is divided into failure and judicial transparency and are described as below:

> *7) Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.*
>
> *8) Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.*

Then in the 9th principle, it makes the designers and builders be responsible for the harms or discrimination caused by their systems and similar to MDRAI, it gives power to the individuals regarding privacy. Also, interestingly, a human-in-the loop concept is put forward in a way that it makes the humans responsible to choose what to delegate to artificial intelligence systems[64].

The last part ends with capability caution[65], the importance of sustainability and thinking about resources[66], risk management[67], supervision of machines such as deep learning[68], and a cautionary message that general artificial intelligence cannot be developed just for the sake of one nation or one organization alone[69].

B. Main Ethical Frameworks

Now that the three (3) ethical framework declarations on artificial intelligence were analyzed, this paper will go through the most important ethical frameworks before analyzing the AIDA.

[60] Asilomar AI Principles Principle 1.

[61] Asilomar AI Principles Principle 3.

[62] Asilomar AI Principles Principle 5.

[63] Asilomar AI Principles Principles 7 & 8.

[64] Asilomar AI Principles Principle 16.

[65] Asilomar AI Principles Principle 19.

[66] Asilomar AI Principles Principle 20.

[67] Asilomar AI Principles Principle 21.

[68] Asilomar AI Principles Principle 22.

[69] Asilomar AI Principles Principle 23.

1. Privacy

Privacy is one of the most important frameworks for artificial intelligence. It is all about respecting individual privacy and protecting it. As mentioned above, this principle is so important that there are already many regulations around the world that cover this principle. In the current era, data replaces oil[70]. As a matter of fact governments and companies massively collect personal information nowadays to either disclose (or sell) to another party or use them for various purposes such as marketing.

In the field of artificial intelligence, data is not just a nice-to-have but rather is a must. The more you have, the better it is for the developers and designers. Artificial intelligence is all about statistics and so the input is as important as the output. So firms will be data hungry and will do anything to collect all kinds of information from individuals and sometimes, the public without even knowing about it. Once collected, they will be used as training data, testing data and validation data, so that ultimately, they could be useful for the creation of different algorithmic models that are used for various purposes. If a firm knows a person's favorite food, allergies, illness history and eating habits, there could be all kinds of strategies to target that person through personalized marketing[71].

Then the next question would be "why do we need protection?". Let's take the Tim Hortons case[72] in Canada to better understand the impact and the importance of protecting our personal information for privacy. As a summary, in May 2019, Tim Hortons started to collect geolocation data of the users that have downloaded the Tim Hortons smartphone application. The collection of geolocations was mentioned to be only functional when a user would open the application based on the FAQ. However, the user's geolocation was actually tracked even when the application was not open. Thus, the application was able to infer a user's home, workplace, school, favorite stores, clinics, and even vacation place. One should remember the Target's case[73] where with inference based from buying patterns (through rebate coupons), Target was able to figure out that a teenage girl was pregnant even before her father knew. Furthermore, let's imagine that Tim Hortons was hacked, even though Tim Hortons confirmed that all personal data would be deleted, before the data was deleted. Personal informations are sold in the dark web easily[74] and this information could be used for various criminal

[70] Javier Fernández-Lasquetty, 'A Data Economy: The Oil of the 21st Century' (*IE University* 19 June 2020), <www.ie.edu/building-resilience/knowledge/data-economy-oil-21st-century/> accessed on August 8, 2022.

[71] Timothy Caulfield, 'The problem with personalized health information' *Policy Options* (9 December 2019), <https://policyoptions.irpp.org/magazines/december-2019/the-problem-with-personalized-health-information/?mc_cid=0fdfb6a1e0&mc_eid=186b1383ed> accessed on August 8, 2022.

[72] Office of the Privacy Commissioner of Canada, 'Joint investigation into location tracking by the Tim Hortons App' (1 June 2022) PIPEDA Findings #2022-001, <www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2022/pipeda-2022-001/> accessed on August 8, 2022.

[73] Kashmir Hill, 'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did' *Forbes* (16 February 2012), <www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=7403a9166686> accessed on August 8, 2022.

[74] Mary Manzi, Geekflare, 'How Much is Your Personal Information Worth on the Dark Web?' (*Geekflare*, 4 March 2022), <https://geekflare.com/personal-data-on-the-dark-web/#:~:text=Buyers%20can%20purchase%20the%20stolen,used%20to%20make%20fraudulent%20transactions.> accessed on August 8, 2022.

acts such as ransom/threat and social engineering for fraud[75]. This paper does not focus on privacy but with the Tim Hortons example, readers would acknowledge that protection of personal information is important for the general public and consumers.

2. Fairness

The second principle is fairness. There is no clear definition of what fairness is yet but some have attempted to analyze them to understand what fairness is in the artificial intelligence industry[76]. Often, what is fair for engineers and data scientists does not really mean the same thing for the end-users, consumers and the society in general[77]. In general, what ethics and law would consider as fairness would be to eliminate discrimination that could be based on biases as an example. So, when an artificial intelligence system must be fair, it must not produce any discrimination in order to be fair. Furthermore, fairness is related to biases as well. Indeed, within the industry of artificial intelligence, the type and quality of data are important and poor data could lead to biased models[78].

Then one can conclude that the designers or the developers of an artificial intelligence system just need to make sure that the data, the model and the result are all fair, meaning that they would be without any bias and discrimination. Nonetheless, this is not an easy task. In fact, it is easy in theory but in practice, it is not so. Let's take the example of Fintechs providing loans and the ultimate question would be as follows: “should we provide credit loans to the applicant?”. This is a typical classification machine learning model where the output is either yes or no. For Fintech or Big Tech firms, most of the time, they will not necessarily have the financial or credit information of the applicant, so often they would make inferences from alternative data such as social network systems (“SNS”) or commercial platforms. What if a training model shows as a result that candidates that post food pictures on their SNS have higher probability to reimburse a loan? If the Fintech firm would apply the model as it is without ethical impact assessment, would we have a fair artificial intelligence system there? What would happen to the elderly group that may have a good credit record but still not use SNS at all?

There was one interesting case with Amazon’s recruiting system based on machine learning[79]. To summarize, since 2015, Amazon has used machine learning to teach itself to choose the best candidates for its recruiting system. The machine was fed

[75] Ravi Sen, ‘Here’s how much your personal information is worth to cybercriminals – and what they do with it’ (*The Conversation,* 13 May, 2021) <https://theconversation.com/heres-how-much-your-personal-information-is-worth-to-cybercriminals-and-what-they-do-with-it-158934> accessed on August 8, 2022.

[76] Sahil Verma & Julia Rubin, ‘Fairness Definitions Explained’ (2018 IEEE/ACM International Workshop on Software Fairness (FairWare), Gothenburg Sweden May 2018) <https://ieeexplore.ieee.org/abstract/document/8452913> accessed on August 8, 2022.

[77] Genevieve Smith with input and feedback from Nitin Kohli & Ishita Rustagi, ‘What does “fairness” mean for machine learning systems?’ (*Center for Equity, Gender & Leadership (EGAL) at Berkeley Haas*, 2020), <https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf> accessed on August 8, 2022.

[78] Tad Simons, ‘Addressing issues of fairness and bias in AI’ *Thomson Reuters* (30 November 2020) <www.thomsonreuters.com/en-us/posts/news-and-media/ai-fairness-bias/> accessed on August 8, 2022.

[79] Jeffrey Dastin, ‘Amazon scraps secret AI recruiting tool that showed bias against women’ *Reuters* (10 October 2018), <www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> accessed on August 8, 2022.

with resumes of mostly men and the result was that the model preferred men over women. Back then, the IT industry was full of men so because the set of data that were fed were not representative of today or of what the society would actually wish, the output was biased too. In order to combat discriminatory practice through artificial intelligence, New York city has put in place regulations, that will be effective starting first of January 2023, that prohibits the use of automated decision-making process for recruiting decisions unless some of the requirements were met[80].

The list of examples can continue and they will not be all covered in this paper. Nonetheless other than financial product and hiring, there are concerns with health care[81] and education[82]

3. Accountability

The next ethical framework is accountability. Accountability is about being responsible and legally could also mean to be liable. For the artificial intelligence industry, in simple terms, it would mean that if an artificial intelligence system goes wrong, who do we blame for reparation, redress, restitution and punishment? Artificial intelligence does not have legal personality yet besides the humanoid, Sophia, from Saudi Arabia that received citizenship in 2017[83].

In order to have a legal personality, it either needs to be a natural human or be a legal person as of today. As a matter of fact, there are few people that are in favor of granting artificial intelligence systems a legal personality[84]. However, it will not be so easy to grant legal personality to artificial intelligence as many barriers exist[85]. To be cautious, governments, scholars and researchers would prefer the accountability to be held by the human beings behind the artificial intelligence and this is especially true for narrow artificial intelligence. Blaming the results of harms and discrimination on the narrow artificial intelligence will not make any sense today as we mostly are surrounded by narrow artificial intelligence.

However, when general artificial intelligence, that could talk, think and act like human beings, arrive, things may have to be a bit different. As proposed and recommended in paragraph 19 of the Asilomar AI Principles[86], "capability caution" must be considered carefully, which basically says that since not all agree on what

[80] New York City of Council, A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools, ch 20 § 20.

[81] Katherine J. Igoe, 'Algorithmic Bias in Health Care Exacerbates Social Inequities — How to Prevent It' (*Harvard T.H. Chan School of Public Health,* 12 March 2021), <www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/> accessed on August 8, 2022.

[82] Andre M. Perry & Nicol Turner Lee, 'AI is coming to schools, and if we're not careful, so will its biases' (*Brookings*, 26 September, 2019), <www.brookings.edu/blog/the-avenue/2019/09/26/ai-is-coming-to-schools-and-if-were-not-careful-so-will-its-biases/> accessed on August 8, 2022.

[83] Heba Kanso, 'Saudi Arabia gave 'citizenship' to a robot named Sophia, and Saudi women aren't amused' *Reuters* (4 November 2017) <https://globalnews.ca/news/3844031/saudi-arabia-robot-citizen-sophia/> accessed on August 8, 2022.

[84] Visa A.J. Kurki, *A Theory of Legal Personhood,* Oxford University Press 2019.

[85] Simon Chesterman, 'Artificial intelligence and the limits of legal personality'(2020) 69/4 International & Comparative Law Quarterly <https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/artificial-intelligence-and-the-limits-of-legal-personality/1859C6E12F75046309C60C150AB31A29> accessed on August 8, 2022.

[86] Asilomar AI Principles Principle 19.

artificial intelligence can do, we should not assume that there is a limit to what artificial intelligence can do. Simply put, our society should not assume that artificial intelligence could never rule over human beings in the future. Hence accountability should be imposed on a human being, and the firms or governments should not deny responsibility by blaming the artificial intelligence systems. In order to do so, a proper governance should be in place where a senior officer, the board members or a specific committee dedicated to artificial intelligence should have the power to veto. In fact, if the senior officer, board members or a committee has no power, then the governance will not be helpful for accountability, but in the worst case, it will be seen as ethic washing.

4. Transparency/Explainability

The fourth ethical framework is actually a pillar for the accountability principle. In order for a developper, designer or user to be held accountable, one must understand how the artificial intelligence system functions exactly. Transparency is about being open to what is being developed, designed and used. It is also about telling how the systems work, how the data are used, collected and disclosed. Ultimately, the end users or those that are affected by the artificial intelligence systems must be able to understand how the artificial intelligence systems impact them. The concept of transparency includes explainability and interpretability.

Then what are explainability and interpretability? Explainability "*is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. It's easy to miss the subtle difference with interpretability, but consider it like this: interpretability is about being able to discern the mechanics without necessarily knowing why. Explainability is being able to quite literally explain what is happening.*"[87] On the other hand, interpretability is "*about the extent to which a cause and effect can be observed within a system. Or, to put it another way, it is the extent to which you are able to predict what is going to happen, given a change in input or algorithmic parameters. It's being able to look at an algorithm and go yep, I can see what's happening here.*"[88]

Thus, in order to hold an organization accountable for the harm or discrimination caused by the artificial intelligence systems, transparency, explainability and interpretability are the basic foundations because of the black box issue. Dino et al. notes that "*(b)lack boxes map user features into a class or a score without explaining why, because the decision model is not comprehensible to stakeholders, even to expert data scientists.*"[89] As an example, in Korea, Naver Corp. ("Naver") was in the center of criticism with its news portal platform[90]. In the past, Naver was known to do news ranking manipulation on its news portal platform where news was suggested to the

[87] Richard Gall, Packt, 'Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI' *KDNuggets* (December 2018), <www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html> accessed on August 8, 2022.

[88] Ibid.

[89] Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. 'Meaningful Explanations of Black Box AI Decision Systems' (Proceedings of the AAAI Conference on Artificial Intelligence July 2019) <https://doi.org/10.1609/aaai.v33i01.33019780> accessed on August 8, 2022.

[90] Kitae Kim & Shin-Il Moon, 'When Algorithmic Transparency Failed: Controversies Over Algorithm-Driven Content Curation in the South Korean Digital Environment' (2021) 65/6 American Behavioral Scientist <https://journals.sagepub.com/doi/10.1177/0002764221989783> accessed on August 8, 2022.

visitors of the Naver portal. Since then, Naver claimed to have implemented a personalized algorithm to correct the situation where it proposes news according to the interest of the visitors. Nevertheless, despite implementing it, suspicion for manipulation continued and Naver was not able to be fully transparent about the sources of its data and the model. In fact, Naver had failed both with accountability and transparency principles.

Nevertheless, full transparency may also cause harm and create distrust in artificial intelligence systems[91]. Being fully transparent could be beneficial to those who wish to play the adversarial game as once you have the whole map of the game, you know how to cheat to get straight to the result[92]. As an example, if a person knows that a female ordering books via amazon and sending gift cards through SNS to more than ten (10) friends will highly augment the chance to get a loan, that person will play the game to fit into the model.

5. Safety and Security

The last ethical framework to explore is about safety and security of an artificial intelligence system. Artificial intelligence systems are supposed to help and assist the well-being of human beings as declared in the MDRAI. It must cause no harm and when used for safety systems, there must be a control to mitigate the risk of causing harm. Think about self-driving vehicles, smartphones, and other Internet of Thing systems that are constantly evolving. They bring beneficence to humanity but at the same time also carry with them safety and security issues. Imagine if a self-driving vehicle was hacked so that it could be manipulated by a criminal to cause an accident[93]. What would happen if smart speakers were hacked and someone could listen to all the conversation in a household[94]? And there are many other more examples for every time a new technology comes out[95].

According to the Asilomar AI principle, "*AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.*"[96] From the concept and design phase to testing, deployment and maintenance phase, systems embedding emerging technology must be checked and challenged. Let's take the example of Facebook with fake news[97]. In fact, Facebook admitted that it had done nothing to prevent the genocide fuelled by fake news on Facebook regarding the

[91] Philipp Schmidt, Felix Biessmann & Timm Teubner, 'Transparency and trust in artificial intelligence systems' (2020) 29/4 Journal of Decision Systems <https://www.tandfonline.com/doi/full/10.1080/12460125.2020.1819094?scroll=top&needAccess=true> accessed on August 8, 2022.

[92] Paul B. de Laat, 'Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?' (2018) 31 Philos. Technol. <https://rdcu.be/cWUof> accessed on August 8, 2022.

[93] Stephen Ornes, 'How to hack a self-driving car' *Physics World* (18 August 2020) <https://physicsworld.com/a/how-to-hack-a-self-driving-car/> accessed on August 8, 2022.

[94] Charlotte Jee, 'Smart speakers can be hijacked by apps that spy on users' *MIT Technology Review* (21 October 2019) <www.technologyreview.com/2019/10/21/330/smart-speakers-can-be-hijacked-by-apps-that-spy-on-users/> accessed on August 8, 2022.

[95] Roman V. Yampolskiy & M. S. Spellchecker, 'Artificial Intelligence Safety and Cybersecurity:a Timeline of AI Failures' <https://arxiv.org/pdf/1610.07997.pdf> accessed on August 8, 2022.

[96] Asilomar AI Principles, Paragraph 6.

[97] Karen Hao, 'How Facebook got addicted to spreading misinformation' (*MIT Technology Review*, 11 March 2021) <www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/> accessed on August 8, 2022.

Rohingya Muslim minority in Myanmar. The goal of the social network platforms is first to identify the interest of the end-users. Then they feed with what people are interested in so that those people stay longer on that platform and that is how they make money through advertising and marketing through artificial intelligence, modeling, etc. If one trusts in a certain political belief, then a social network platform will feed with similar information as much as possible, be it fake news, so that you stay as much as possible and artificial intelligence, without a proper control, will make it worse[98].

Again, safety and security is all about non-maleficence of artificial intelligence. The developers and designers of artificial intelligence systems should not only check foreseeable harm but use their creativity and imagination since based on capability caution, anything could happen in the future.

III. ANALYSIS OF ARTIFICIAL INTELLIGENCE AND DATA ACT

This paper will now take the AIDA and verify if the five ethical frameworks were taken into consideration at the current form. It will start with the principle of privacy, then fairness, accountability, transparency/explainability and safety & security.

A. Privacy

It seems like AIDA itself does not cover directly the subject of privacy similar to EU's AIA, US's AAA and the Directive. In the EU, the majority of the privacy principle is rather covered in the GDPR, nevertheless, a small section exists in the article 55 of AIA[99] in relation to the regulatory sandbox and article 60 of AIA in relation to the EU database[100]. In the United States, there is no federal legislation regarding privacy, but there are state regulations such as CCPA and in the public sector of Canada, there is the Privacy Act. Therefore, it is likely that AIDA will not have a specific section on privacy with the hope that the CPPA becomes effective.

However, there is indirectly a part concerning privacy in section 6 of AIDA. Section 6 says:

> ***Anonymized data***
>
> *6 A person who carries out any regulated activity and who processes or makes available for use anonymized data in the course of that activity must, in accordance with the regulations, establish measures with respect to*
>
> *(a) the manner in which data is anonymized; and*
>
> *(b) the use or management of anonymized data.*

This section is probably to make sure that developers and designers of artificial intelligence systems cannot use the term anonymization to escape from PIPEDA (or

[98] Gul, A., Erturk, Y. and Elmer, P. (ed.) 'Digital Transformation in Media and Society' (Istanbul University Press 2020), ch 9, Andreas Kaplan 'artificial intelligence, social media, and fake news:is this the end of democracy'.

[99] Artificial Intelligence Act, Article 65.

[100] Artificial Intelligence Act, Article 60.

CPPA). The future regulations will probably provide the requirements for what is considered an anonymized data as well.

B. Fairness

Fairness is about bias and discrimination. In AIDA, section 5 (1) provides definitions on what is a "*biased output*". The definition of "*biased output*" considers the human rights aspect recommended by the TD. Then in section 8 of AIDA, there is a requirement for high-impact systems to implement risk management systems to properly deal with biased output[101]. It looks like that the future regulations will include standards and requirements on how to implement risk management systems that will identify, assess and mitigate risks related to biased output.

Furthermore section 9 of AIDA requires developers and designers to monitor compliance with the mitigating measures (controls). Again, the AIDA does not specify directly what needs to be implemented for a responsible person but future regulations would probably provide more details. Nonetheless we can assume that permanent controls, independent testing, regulatory watch, regular update of the controls library evaluation and update of the controls would be required.

AIDA, compared to AIA, seems to be too simple in this matter. Again, the upcoming regulations should probably add more clarity but some mechanism such as the "human-in-the-loop" requirement for a high-impact system should have been included in the AIDA.

C. Accountability

As for accountability, it first defines in section 5 (2) of AIDA on what a person responsible is:

"*For the purposes of this Part, a person is responsible for an artificial intelligence system, including a high-impact system, if, in the course of international or interprovincial trade and commerce, they design, develop or make available for use the artificial intelligence system or manage its operation.*" So in AIDA, a person, which includes a legal entity, becomes responsible for the artificial intelligence system as a designer, developer or for making available for use in the international or interprovincial trade and commerce. The concept of developer and designer are limited but the wording "make available for use" is broad enough to capture different players as well.

Then AIDA allocates accountability to the person responsible to 1) put in place risk management program[102], 2) monitor mitigating measures[103], 3) keep records[104] and 4) disclosing regulatory information[105]. Furthermore, a person who violates the AIDA requirements could be faced both to administrative monetary penalties[106]("AMP") and

[101] Artificial Intelligence and Data Act, section 2.
[102] Artificial Intelligence and Data Act, section 8.
[103] Artificial Intelligence and Data Act, section 9.
[104] Artificial Intelligence and Data Act, section 10.
[105] Artificial Intelligence and Data Act, sections 11 & 12.
[106] Artificial Intelligence and Data Act, section 29.

offences[107]. For AMP, more information will be provided in the regulations. As for the offences, any person who violates sections 6 to 12 and mislead or provide false information are all guilty of an offence"[108]. Similar to AIDA and GDPR, the maximum amount for a conviction on indictment could go up to $10,000,000 or 3% of the company's gross global revenues, whichever is greater.

In the EU, AIA on the other hand defines "providers", "users", "importers", "distributors" and "operators". In Chapter 3 of AIA, it provides all the obligations of the providers, users and other parties. Very prescriptive obligations are written down so that each actor understands what they are accountable for[109]. To add, similar to AIDA, there are penalties mechanisms within AIA as well[110].

Accountancy also means having a strong and independent organization with power to strengthen the compliance and AIDA mentions about an "Artificial Intelligence and Data Commissioner[111]". A commissioner without any enforcing power like the current Office of Privacy of Commissioner ("OPC") of Canada should be avoided[112]. Furthermore, the public would prefer to have more of an independent regulator than a commissioner chosen directly from the Minister. The public sector could be itself biased and may not be the best to play such a role. Moreover, the private sectors would want an ombudsman type like the OPC[113] with more power instead of having a public agency. In addition, the audit system is a bit unclear as it seems to let the person responsible of high-impact system choose its own auditor and it is really in the hands of the Minister to figure out that something is not going on well in the industry[114]. In the end, to let all the power in the hands of the public sector may be a real concern as the government itself acknowledges[115] that there is a systemic risk at the government level and that the government is also trying to combat discrimination [116].

D. Transparency/Explainability

For transparency and explainability ethical frameworks, AIDA puts several requirements on the designers & developers as well as for those who manage high-impact systems.

[107] Artificial Intelligence and Data Act, section 30.

[108] Artificial Intelligence and Data Act, section 30 (1) & (2).

[109] Artificial Intelligence Act, Articles 16 to 29.

[110] Artificial Intelligence Act, Articles 71 and 72.

[111] Artificial Intelligence and Data Act, section 33.

[112] Barry Sookman, 'PIPEDA by the numbers: lessons for privacy law reform in Canada?' (*Barry Sookman*, 12 October 2020)<www.barrysookman.com/2020/10/12/pipeda-by-the-numbers-lessons-for-law-reform/> accessed on August 8, 2022.

[113] To note that if the proposed CPPA becomes effective, the OPC will likely have more power than now.

[114] Artificial Intelligence and Data Act, section 15.

[115] 'Statement by the Prime Minister on the International Day for the Elimination of Racial Discrimination' (*Prime Minister of Canada*, 21 March 2022) <https://pm.gc.ca/en/news/statements/2022/03/21/statement-prime-minister-international-day-elimination-racial> Accessed on August 8, 2022.

[116] 'Addressing Systemic Racism in Canada's Criminal Justice System while maintaining public safety: proposed legislative amendments to the Criminal Code and the Controlled Drugs and Substances Act' (*Government of Canada*, 7 December 2021) <www.justice.gc.ca/eng/csj-sjc/pl/sr-rs/index.html> accessed on August 8, 2022.

Designers and developers must "*publish on a publicly available website a plain-language description of the system that includes an explanation of*

(a) how the system is intended to be used;

(b) the types of content that it is intended to generate and the decisions, recommendations or predictions that it is intended to make;

(c) the mitigation measures established under section 8 in respect of it; and

(d) any other information that may be prescribed by regulation."[117]

As for those who manages, they must "*publish on a publicly available website a plain-language description of the system that includes an explanation of*

(a) how the system is used;

(b) the types of content that it generates and the decisions, recommendations or predictions that it makes;

(c) the mitigation measures established under section 8 in respect of it; and

(d) any other information that may be prescribed by regulation.[118]"

Furthermore, a person must keep general records[119] and those who are responsible for high-impact systems must notify the Minister if the system will likely result or results in material harm[120]. However, what "*material harm*" means is not defined in AIDA and unless it is specified in the upcoming regulations, unfortunately it should be expected that the definition would be developed through cases. Also, guidelines could be provided to give examples of what could be considered a material harm to the industry.

In addition, transparency principles related to automated decision-making systems in CPPA and Quebec's Bill 64[121] will normally provide more transparency. As mentioned above, the definition of an automated decision-making system is broader and covers artificial intelligence as well. In CPPA, an organization must be transparent about the use of automated decision-making system[122] and when used for an individual that could have a significant impact, the individual has the right to request in writing for explanation of the prediction, recommendation and decision[123]. Moreover, it requires the organization to include in the explanation information such as "*the type of personal information that was used to make the prediction, recommendation or*

[117] Artificial Intelligence and Data Act, section 11 (1).

[118] Artificial Intelligence and Data Act, section 11 (2).

[119] Artificial Intelligence and Data Act, section 10.

[120] Artificial Intelligence and Data Act, section 12.

[121] Bill 64, *An Act to modernize legislative provisions as regards the protection of personal information*, 1st session, 42nd Legislature, 2021.

[122] Consumer Privacy Protection Act, section 62 (2) (c).

[123] Consumer Privacy Protection Act, section 63 (3).

decision, the source of the information and the reasons or principal factors that led to the prediction, recommendation or decision."[124]

Similarly, in Quebec, section 12.1 requires organizations to communicate to individuals when an automated decision-making system is exclusively used and when a request for explanation is made, the organizations must provide the rationale why certain results came out from the automated decision-making system. Besides, individuals seem to have more power than the CPPA as they can require a human to review the automatically made decision. However, when compared with article 22 of the GDPR, Quebec provides less power as in the EU, individuals have the power to force the organization not to solely use automated decision making systems to make decisions[125].

E. Safety and Security

Finally, AIDA uses the concept of "harm" to grasp the safety and security ethical framework. As previously noted above, harm includes physical and psychological harm to individuals. A responsible person must perform risk management for harms that could be caused by a high-impact system[126] and monitor the controls put in place to mitigate the risks identified in the risk assessment process[127]. Moreover, as mentioned in the transparency/explanation section, should there be a material harm caused by a high-impact system, the person responsible must notify the Minister[128]. Additionally, the Minister has the power to request for more information[129] and order the person responsible to cease using or making available the high-impact system[130]. The minister may also publish on a public website information about an artificial intelligence system that may cause serious risk of imminent harm[131].

As for punishment, any person who makes available for use which causes serious physical or psychological harm will be considered to have committed an offence[132] and in consequence will be subject to section 40 of AIDA for penalties. For conviction on indictment of an individual, it will either be a fine or/and imprisonment of up to five (5) years and for other persons, it will be a maximum fine of $25,000,000 or four (4) % of global gross revenue, whichever is greater.

Also, it is important to note that CCPA[133] also has sections covering harms in relation to the personal information that may provide additional obligations to the person responsible for artificial intelligence.

CONCLUSION

The goal of this paper was to take AIDA and analyse based on the five main ethical frameworks of artificial intelligence. AIDA is still a draft and its regulations are

[124] Consumer Privacy Protection Act, section 63 (4).
[125] GDPR, Article 22 (1).
[126] Artificial Intelligence and Data Act, section 8.
[127] Artificial Intelligence and Data Act, section 9.
[128] Artificial Intelligence and Data Act, section 10.
[129] Artificial Intelligence and Data Act, section 14.
[130] Artificial Intelligence and Data Act, section 17.
[131] Artificial Intelligence and Data Act, section 28.
[132] Artificial Intelligence and Data Act, section 39.
[133] Consumer Privacy Protection Act, sections 58 & 59.

yet to be published so this paper cannot confirm the overall ethical framework of the federal legislation. However, this paper was able to confirm that AIDA covers the basics of five main artificial intelligence ethical frameworks, besides privacy principle which is covered by PIPEDA just like in the EU.

However, AIDA has room for improvements as well. First of all, there seems to be an issue with the scope. It does not apply to the public sector. The Directive is not perfect and has its own issues[134] as it does not cover candidates and employees. Thus, the public sector employees will have neither the Directive nor AIDA applicable to them. Another big issue is the transparency with audit results. Also, the RCMP case with Clearview AI is a good example of why public sectors should also be covered, at least at the federal level and let the provincial governments adopt it in a similar way.

Next, Although AIDA covers ethical principles for artificial intelligence such as fairness, accountability, transparency/explainability and safety & security, it seems to lack clarity. As an example, it will be hard for the industry to know that a "*material harm*" is. Nevertheless, guidelines could be created to provide examples of what material harm is in the future for the industry. Moreover, by comparison with AIA, which is more prescriptive, many important aspects of ethical artificial intelligence principles are missing. As an example, article 14 of AIA provides the human oversight requirements in the EU[135]. Another example could be the article 15 about *Accuracy, robustness and cybersecurity* where for high-risk AI systems, there are additional requirements around accuracy, robustness and cybersecurity[136]. Once the regulations become publicly available, it would be clearer for the industry but the regulations should have similar items that are reflected in AIA as AIDA does not include them in the current version.

Last, to have real accountability, there should be an independent and neutral agency that can be trusted by the states, the public and private sector actors. In the current version of AIDA, persons responsible for artificial intelligence systems must perform the assessment mentioned in section 7 of AIDA[137]. This would probably provide a lot of freedom to the private sectors but at the same time be ambiguous with lack of oversight. As previously mentioned, the public sector itself may be biased so to let the Minister and commissioner play alone the role of the judge would be dangerous. This paper suggests that to have a real accountability implemented in Canada, It might be better to have an independent organization that would include experts like data scientists, artificial intelligence engineers, lawyers, medical experts, psychologists, sociologists, human resource experts, public affairs experts, and etc.

[134] Omar Bitar, Benoit Deshaies & Dawn Hall, '3rd Review of the Treasury Board Directive on Automated Decision-Making' (29 April 2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4087546> accessed on August 8, 2022.

[135] Artificial Intelligence Act, Article 14.

[136] Artificial Intelligence Act, Article 15.

[137] Artificial Intelligence and Data Act, section 7.